R-AGT: Agent Rules

Agent Rules

AI agents are first-class participants in this framework. These rules ensure agents are accountable, auditable, and structurally cannot exceed the permissions of the humans who delegate to them.

R-AGT-01 All Agent Actions Logged as PROV-O

Every action taken by an AI agent — read, write, policy evaluation, credential check — must produce a W3C PROV-O provenance record. The record must name: (a) the agent DID, (b) the delegating human DID, (c) the action type, (d) the entity acted upon, (e) timestamp.

Implementation: PROV-O records written to `provenance_log` append-only table synchronously with the action. Failed PROV-O write aborts the action transaction.

R-AGT-02 Policy Check Mandatory Before Every Write

An agent MUST call `POST /policy/evaluate` before any write operation (submission, validation signature, credential issuance, ledger entry). The policy evaluation result must be `allow: true` before the write proceeds. This is enforced at the API gateway level — agents cannot bypass it.

Implementation: API gateway middleware intercepts all write requests. `X-Policy-Token` header must contain a fresh (< 30s) OPA evaluation token. Requests without valid token receive 403.

R-AGT-03 FPIC Block is Absolute

When FPIC is not granted or has been revoked for a territory, no agent action may read or write data for that territory. This is not an OPA rule — it is enforced at Postgres row-level security, below the API layer. There is no override, no emergency bypass, no admin exception.

Implementation: Postgres RLS policy: `fpic_status = 'active'` required on every row. Even CTH Steward credentials do not bypass this. The only way to lift the block is for the Community Sovereign to re-grant FPIC.

?? Agent design principle: an agent inherits the delegating human's permissions — never more. If the human cannot do something, neither can the agent acting on their behalf. Credential delegation is explicit and recorded in the JWT claims.

Revisión #1
Creado 2026-05-27 13:47:38 UTC por Angelica Diaz
Actualizado 2026-05-27 13:47:38 UTC por Angelica Diaz